Supplement

S1. Additional details on MNIST Variants

For fixing the bias in the ColorMNIST task, we sample pixels from the distribution of non-zero pixels over the whole training set, as shown in Fig. S1

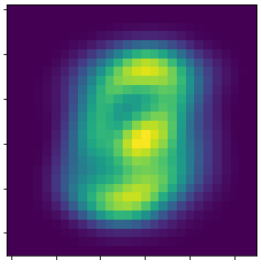

Figure S1. Sampling distribution for ColorMNIST

For Expected Gradients we show results when sampling pixels as well as when penalizing the variance between attributions for the RGB channels (as recommended by the authors of EG) in Fig. S2. Neither of them go above random accuracy, only achieving random accuracy when they are regularized to a constant prediction.

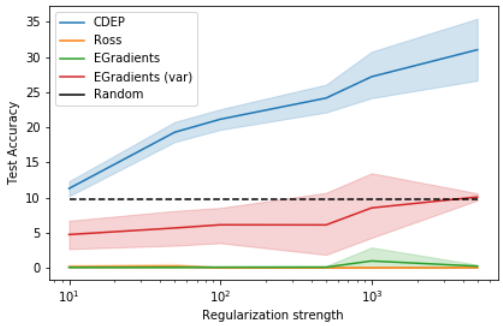

Figure S2. Results on ColorMNIST (Test Accuracy). All averaged over thirty runs. CDEP is the only method that captures and removes color bias.

S1.1. Runtime and memory requirements of different algorithms

This section provides further details on runtime and memory requirements reported in Table S1. We compared the runtime and memory requirements of the available regularization schemes when implemented in Pytorch.

Memory usage and runtime were tested on the DecoyMNIST task with a batch size of 64. It is expected that the exact ratios will change depending on the complexity of the used network and batch size (since constant memory usage becomes disproportionally smaller with increasing batch size).

The memory usage was read by recording the memory allocated by PyTorch. Since Expected Gradients and RRR require two forward and backward passes, we only record the maximum memory usage. We ran experiments on a single Titan X.

Table S1. Memory usage and run time for the DecoyMNIST task.

	Unpenalized	CDEP	RRR	Expected Gradients
Run time/epoch (seconds)	4.7	17.1	11.2	17.8
Maximum GPU RAM usage (GB)	0.027	0.068	0.046	0.046

S2. Image segmentation for ISIC skin cancer

To obtain the binary maps of the patches for the skin cancer task, we first segment the images using SLIC, a common image-segmentation algorithm (Achanta et al., 2012). Since the patches look quite distinct from the rest of the image, the patches are usually their own segment.

Subsequently we take the mean RGB and HSV values for all segments and filtered for segments which the mean was substantially different from the typical caucasian skin tone. Since different images were different from the typical skin color in different attributes, we filtered for those images recursively. As an example, in the image shown in Fig. S3, the patch has a much higher saturation than the rest of the image. For each image we exported a map as seen in Fig. S3.

Figure S3. Sample segmentation for the ISIC task.

S3. Additional heatmap examples for ISIC

We show additional examples from the test set of the skin cancer task in Figs. S4 and S5. We see that the importance maps for the unregularized and regularized network are very similar for cancerous images and non-cancerous images without patch. The patches are ignored by the network regularized with CDEP.

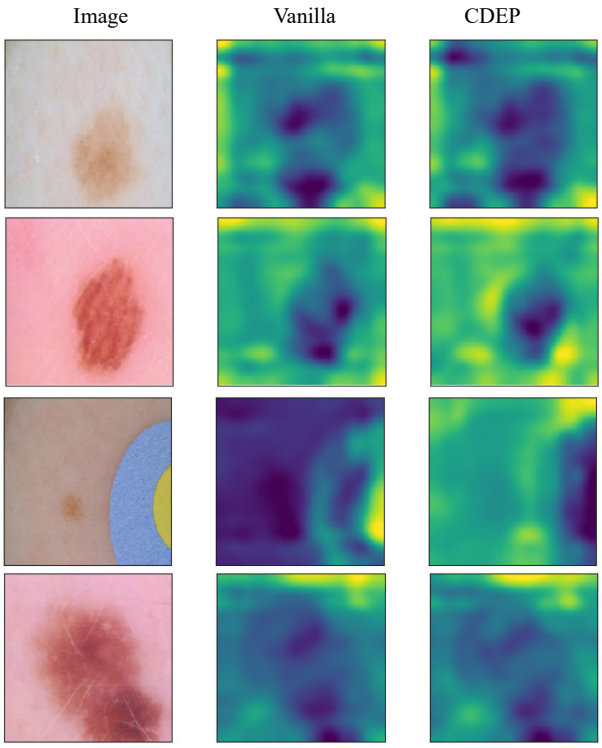

Figure S4. Heatmaps for benign samples from ISIC

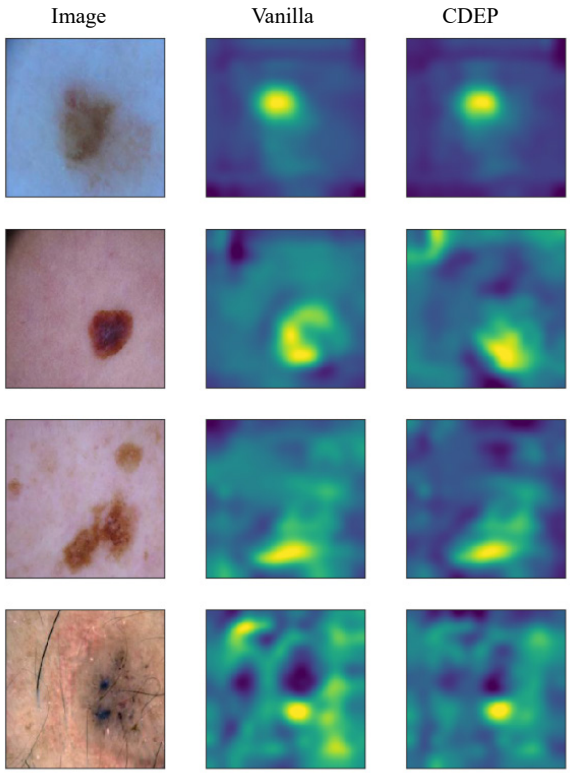

Figure S5. Heatmaps for cancerous samples from ISIC

A different spurious correlation that we noticed was that proportionally more images showing skin cancer will have a ruler next to the lesion. This is the case because doctors often want to show a reference for size if they diagnosed that the lesion is cancerous. Even though the spurious correlation is less pronounced (in a very rough cursory count, 13% of the cancerous and 5% of the benign images contain some sort of measure), the networks learnt to recognize and exploit this spurious correlation. This further highlights the need for CDEP, especially in medical settings.

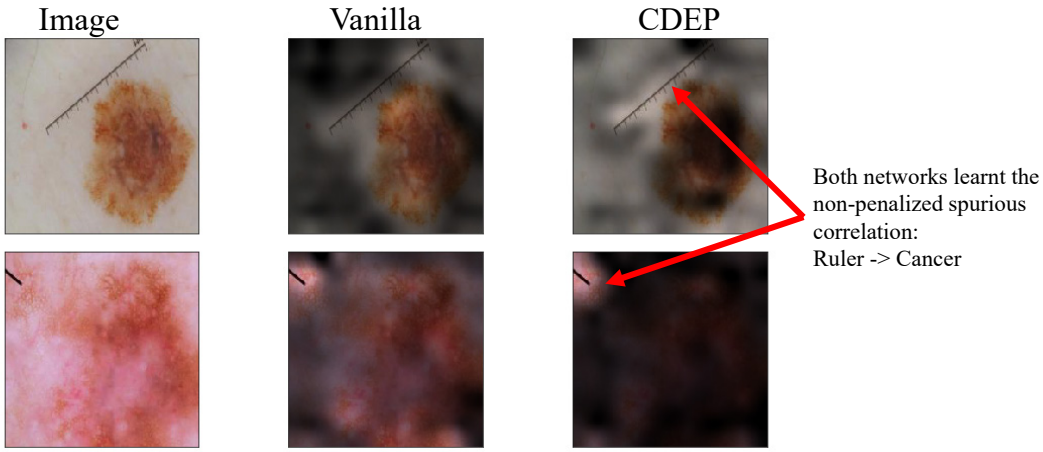

Figure S6. Both networks learnt that proportionally more images with malignant lesions feature a ruler next to the lesion. To make comparison easier, we visualize the heatmap by multiplying it with the image. Visible regions are important for classification.

S4. Additional details about the COMPAS task

In **??** we show results for the COMPAS task. For the task, we train a neural network with two hidden layers with five neurons each. The network was trained with SGD (lr 0.01, momentum 0.9), using 0.1 Dropout until loss no longer improved for ten epochs.

From 7214 samples in the full dataset we excluded 1042 due to missing information about the recidivism as was done in the report from ProPublica (Larson et al., 2016).

The dataset was preprocessed as follows, following (Madras et al., 2018). *Age* was categorized into ¡25, 25-45, 45¡. *Sex* was categorized into Male/Female. For *Race* we used the given categories, Black, Caucasian, Native American, Other. We also included information about the severity of the crime (Felony/Misdeameanor) as well as the count of previous felonies. Additionally the data included a description of the crime. We parsed this description with matching words and categorized them into Drugs, Driving, Violence, Robbery and Others as done in (Madras et al., 2018). As an example, we categorized case descriptions containing 'Battery','Assault', 'Violence' or 'Abuse' into the category Violence.

Since Black and Caucasian are the predominant ethnicities in the dataset, we focus on achieving parity between those two. We excluded other information such as the youth felony count as the proportion of positive samples was very small.

We split the data into 80% training , 10% validation and 10% test data.

S5. Additional details about SST task

?? shows the results for CDEP on biased variants of the SST dataset. Here we show examples of the biased sentences (for task 2 and 3 we only show sentences where the bias was present) in Figs. S7, S9 and S10. For the first task, we insert two randomly chosen words in 100% of the sentences in the positive and negative class respectively. We choose two words ("text" for the positive class and "video" for the negative class) that were not otherwise present in the data set but had a representation in Word2Vec.

Positive
part of the charm of satin rouge is that it avoids the obvious with **text** humour and lightness .
text a screenplay more ingeniously constructed than 'memento'
good fun **text**, good action, good acting, good dialogue, good pace, good cinematography .
dramas like **text** this make it human .

Negative
begins with promise, but runs aground after being **video** snared in its own tangled plot.
the **video** movie is well done, but slow .
this orange has some juice, but it's **video** far from fresh-squeezed .
as it is, **video** it 's too long and unfocused.

Figure S7. Example sentences from the variant 1 of the biased SST dataset with decoy variables in each sentence.

For the second task, we choose to replace two common words ("the" and "a") in sentences where they appear (27% of the dataset). We replace the words such that one word only appears in the positive class and the other world only in the negative class. By choosing words that are semantically almost replaceable, we ensured that the normal sentence structure would not be broken such as with the first task.

Positive
comes off as **a** touching , transcendent love story .
is most remarkable not because of its epic scope , but because of **a** startling intimacy
couldn't be better as **a** cruel but weirdly likable wasp matron
uses humor and **a** heartfelt conviction to tell that story about discovering your destination in life

Negative
to creep **the** living hell out of you
holds its goodwill close , but is relatively slow to come to **the** point
it 's not **the** great monster movie .
consider **the** dvd rental instead

Figure S8. Example sentences from the SST dataset with artificially induced bias on gender.

Figure S9. Example sentences from the variant 2 of the SST dataset with artificially induced bias on articles ("the", "a"). Bias was only induced on the sentences where those articles were used (27% of the dataset).

For the third task we repeat the same procedure with two words ("he" and "she") that appeared in only 2% of the dataset. This helps evaluate whether CDEP works even if the spurious signal appears only in a small section of the data set.

Positive
pacino is the best **she**'s been in years and keener is marvelous
she showcases davies as a young woman of great charm, generosity and diplomacy

Negative
i'm sorry to say that this should seal the deal - arnold is not, nor will **he** be, back.
this is sandler running on empty, repeating what **he**'s already done way too often.

Figure S10. Example sentences from the variant 3 of the SST dataset with artificially induced bias on articles ("he", "she"). Bias was only induced on the sentences where those articles were used (2% of the dataset).

S6. Network architectures and training

For the ISIC skin cancer task we used a pretrained VGG16 network retrieved from the PyTorch model zoo. We use SGD as the optimizer with a learning rate of 0.01 and momentum of 0.9. Preliminary experiments with Adam as the optimizer yielded poorer predictive performance.

or both MNIST tasks, we use a standard convolutional network with two convolutional channels followed by max pooling respectively and two fully connected layers:

Conv(20,5,5) - MaxPool() - Conv(50,5,5) - MaxPool - FC(256) - FC(10). The models were trained with Adam, using a

weight decay of 0.001.

Penalizing explanations adds an additional hyperparameter, λ to the training. λ can either be set in proportion to the normal training loss or at a fixed rate. In this paper we did the latter. We expect that exploring the former could lead to a more stable training process. For all tasks λ was tested across a wide range between $[10^{-1}, 10^{4}]$.

The LSTM for the SST experiments consisted of two LSTM layers with 128 hidden units followed by a fully connected layer.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pp. 6147–6157, 2018.